

Quantitative Comparison of Pre-explosions and Subheadings with Methodologic Search Terms in MEDLINE

NL Wilczynski, CJ Walker, KA McKibbin, RB Haynes

Health Information Research Unit,

Dept. of Clinical Epidemiology & Biostatistics,

McMaster University, HSC, Room 3H7,

1200 Main St W, Hamilton, Ont, Canada L8N 3Z5

(905)525-9140 x22311, FAX 905-546-0401, E-MAIL WILCZYN@McMASTER.CA

ABSTRACT

Objective: To compare the retrieval characteristics of subheadings with methodologic textwords and MeSH terms in MEDLINE for identifying sound clinical studies on the etiology, prognosis, diagnosis, prevention and treatment of disorders in general adult medicine.

Design: Analytic survey of the information retrieval properties of methodologic textwords, single methodologic MeSH terms, pre-explosions and subheadings selected to detect studies meeting basic methodologic criteria for direct clinical use in general adult medicine.

Measures: The sensitivity, specificity, and precision of search terms were determined by comparing the citations retrieved by the search strategies in MEDLINE with that of a manual review (the gold standard) of all articles in 10 internal and general medicine journals for 1986 and 1991.

Results: For treatment and diagnosis in 1991, and treatment, diagnosis, and etiology in 1986, pre-explosions yielded the highest sensitivity, with typical absolute increases exceeding 15%. For etiology and prognosis in 1991, and prognosis in 1986, textwords or MeSH terms yielded the highest sensitivity. In all cases the increase in sensitivity was coupled with a loss in specificity and precision.

Conclusions: Compared with searching with single methodologic textwords and subject headings, the detection of sound clinical studies on the diagnosis and treatment of disorders in general adult medicine was consistently enhanced by searching with pre-explosions, but at a price of decreased specificity and precision.

INTRODUCTION

It is important for clinical end users of MEDLINE to be able to retrieve articles that are both scientifically sound and directly relevant to clinical practice. MEDLINE, however, is a general purpose biomedical research literature database, with only a small proportion of articles reporting evidence that can be directly applied in clinical practice. A potential method for improving the detection of studies of high quality for clinical practice is the use of "methodologic

search filters" [1]. A methodologic search filter is a search term or terms (such as 'random allocation' for sound studies of medical intervention) that select studies that are at the most advanced stages of testing for clinical application. The retrieval performance, however, of such terms on search recall and precision has not been fully tested. The purpose of this study was to test individual methodologic Medical Subject Headings (MeSH) terms and textwords in common use, and permutations and combinations of these MeSH terms and textwords for identifying studies meeting basic methodologic criteria on the etiology, prognosis, diagnosis, prevention and treatment of disorders in general adult medicine. In this paper, the information retrieval properties of subheadings are compared with textwords and MeSH terms. The retrieval properties of individual textwords and MeSH terms were reported previously [2]. Our results are of most interest to clinicians doing their own searches for clinically relevant and valid studies and for librarians involved in assisting clinicians to construct their own searches.

METHODS

The study compared the retrieval performance of methodologic search terms, pre-explosions, and subheadings in MEDLINE with a manual review of each article for each issue of 10 internal and general medicine journals for the 2 years 1986 and 1991. To evaluate MEDLINE strategies designed to retrieve studies meeting basic methodologic criteria for clinical practice, terms related to research design features were run as search strategies and treated as "diagnostic tests" for sound studies as determined by the manual review of the literature, treated as the "gold standard". Borrowing from the concepts of diagnostic test evaluation and library science, the sensitivity, specificity, and precision of MEDLINE searches were determined. The sensitivity of the MEDLINE search strategies was calculated as the proportion of correctly detected citations with relevant content and sound study methods among all relevant citations as defined by the manual review of the literature. This is equivalent to the library term 'recall'. Specificity was the proportion of

irrelevant, unsound studies excluded by the search strategy. This differs from precision which is the proportion of all articles retrieved by a search strategy that are sound and relevant.

Manual Review of the Literature

For the years 1986 and 1991, 3 research assistants hand searched 10 journals, the same 10 in each year, for studies meeting methodologic criteria on the etiology, prognosis, diagnosis, prevention and treatment of human adult disease. The 10 journals searched were *American Journal of Medicine*, *Annals of Internal Medicine*, *Archives of Internal Medicine*, *BMJ* (*British Medical Journal* in 1986), *Circulation*, *Diabetes Care*, *Journal of Internal Medicine* (*Acta Medica Scandinavica* in 1986), *Journal of the American Medical Association*, *The Lancet*, and *New England Journal of Medicine*, including supplements. These journals were selected on the basis of impact factors and immediacy indexes [3], and to provide a broad range of publications, including both internal and general medical journals and both American and European authors.

Articles were classified for 'format', 'interest', 'purpose' and 'methodologic rigor'. 'Format' categories included 'original study', 'review', 'general article', 'conference report', 'decision analysis', and 'case report'. Articles with more than one format were classified for all that applied. An 'original study' was defined as any full text article in which the investigators had made first-hand observations. A 'review' was any full text article that was bannered review, had review in the title or in a section heading, or indicated in the text that the intention was to review or summarize the literature on a topic. A 'general article' was a general or philosophical discussion of a topic without original first-hand observation or a statement that the purpose was to review or appraise a body of knowledge, including unbannered news items, unbannered editorials, position and opinion papers, musings and psychosocial observations. A 'conference report' was defined as such by the journal but was reclassified by us as an original or review article when meeting those criteria. A 'decision analysis' was defined as the breaking down of the management of patients into component parts, defining routes of management and consequences of management based on alternatives, for the purpose of defining optimal methods of management. A 'case report' was defined as an original study involving less than 10 subjects. Items excluded from classification included bannered letters to the editor, book reviews, announcements, policy watch, editorials, commentaries, brief clinical observations, correspondence, news, obituaries, postgraduate and continuing education

forums, and notices.

To be considered of 'interest' to the medical care of human adults the study had to be concerned with the understanding and management of clinical problems with clinical endpoints and recommendations for applications in human subjects, at least 50% of whom were ≥ 18 years of age at study entry. All format categories were classified for interest.

Articles classified as original studies, reviews, or case reports and of interest were classified for 'purpose'. Articles could have more than one purpose and were classified for all that applied. Articles were classified as 'etiology' when the content pertained directly to causation of a disease or condition; as 'prognosis' when the content pertained directly to the prediction of the clinical course or the natural history of a disease with the disease existing at the beginning of the study; as 'diagnosis' when the content pertained directly to the evaluation of a disease process, usually through comparing methods of arriving at a diagnosis; as 'treatment or prevention' when the content pertained directly to therapy, prevention or rehabilitation; and as 'something else' when the purpose of the study was something other than the above.

Studies in each purpose category were evaluated for 'methodologic rigor' and were assessed to determine if they met one key methodologic criterion specific to their purpose as shown in Table 1. These criteria were based on critical appraisal criteria for applied research [4] but were set at a minimal level in recognition that few published studies meet the full set of criteria for unbiased clinical evaluation.

Table 1. Key methodologic criteria by purpose of study

Purpose	Key methodologic criterion
Etiology	Formal control group: random or quasi-random allocation of participants to treatment and control groups; or the study was a non-randomized, concurrent control trial, a cohort analytic study with matching or statistical adjustment to create comparable groups, or a case-control study
Prognosis	A cohort of subjects all having the disease in question at baseline without the outcome of interest
Diagnosis	Provision of sufficient data to calculate the sensitivity and specificity of the test or likelihood ratios based on subjects who had all been tested on both the test and diagnostic standard
Treatment	Random or quasi-random allocation of participants to treatment and control groups
Review	Reproducible description of the methods for conducting the review

Inter-rater reliability was assessed for the classification of articles for format, interest, purpose and methods. In all cases the degree of agreement beyond chance was assessed by the kappa statistic and was greater than 0.80.

The sample size required to detect a 20% improvement in sensitivity for the comparison of one MEDLINE search strategy with another on the same topic was 73 methodologically sound studies in each of the purpose categories for each of the years 1986 and 1991 (type 1 error of 5%, one-sided, and a type 2 error rate of 20%).

Collecting Search Terms

To construct a comprehensive set of search terms, we began a list of methodologic subject headings and textwords and then sought input from clinicians and librarians in the United States and Canada through interviews of known searchers; requests on several electronic bulletin boards and in national publications, meetings and conferences; and requests to the National Library of Medicine and Canada Institute for Scientific and Technical Information. Individuals were asked what terms or phrases they used when searching for studies of etiology, prognosis, diagnosis, prevention and treatment and related review articles. Terms could be from MeSH, including publication types, check tags, pre-exposures (subheading pre-exposure groups together and retrieves subheadings that relate to the particular clinical category being studied; e.g. the subheading pre-exposure therapeutic use includes the subheadings administration & dosage, adverse effects, contraindication, and poisoning in addition to the subheading therapeutic use) and subheadings, or textwords denoting applied research methodology in titles and abstracts of articles. The list, excluding inaccurate terms, appears in the Appendix. Some of the terms and phrases were different for the 2 years as some of the corresponding terms changed definitions and some terms retrieved 0 citations for the 10 journals in 1986 and/or 1991.

DATA COLLECTION

Manual ratings of articles in the 10 journals for 1986 and 1991 were recorded on data collection forms, and the bibliographic information, including the 8-digit unique identifier, for the articles in those journals was captured from MEDLINE. Each journal title was searched in MEDLINE for 1986 and 1991 and the publication types 'editorial,' 'comment,' 'letter' and 'news' were eliminated from the search using the boolean 'AND NOT' operator.

The MeSH terms and textwords to be tested were searched in MEDLINE for 1986 and 1991 for the

10 journals. The unique identifiers were captured and then linked with the manual review data.

TESTING STRATEGIES

All methods terms were tested, both individually and in combination, and the sensitivity, specificity, and precision was calculated. For 1991 there were 27 etiology terms, 28 prognosis terms, 25 diagnosis terms, and 26 treatment terms. For 1986 there were 20 etiology terms, 22 prognosis terms, 25 diagnosis terms, and 20 treatment terms (see Appendix).

RESULTS

The results of the manual review of the journals was previously reported [2]. Briefly, the total number of original, review and case report articles in 1991 was 3495, and in 1986 was 3682. Less than half of the studies cited met basic criteria for scientific merit for clinical application.

For 1991, the sensitivity, specificity, and precision of the single best terms and subheadings are presented in Table 2. The corresponding figures for 1986 are presented in Table 3.

Table 2. Sensitivity, specificity and precision for single best terms and subheadings in 1991

Category	Search strategy	Sensitivity	Specificity	Precision
1991				
Etiology	Risk (tw) (best single term)	0.67	0.79	0.15
	Etiology & (px)	0.63	0.56	0.07
	Etiology (sh)	0.40	0.78	0.09
Prognosis	Exp Cohort Studies (best single term)	0.60	0.80	0.11
	Mortality (sh)	0.53	0.93	0.20
Diagnosis	Sensitivity (tw) (best single term)	0.57	0.97	0.33
	Diagnosis & (px)	0.80	0.77	0.09
	Diagnosis (sh)	0.59	0.88	0.13
	Diagnostic use (sh)	0.26	0.96	0.18
Treatment	Clinical Trial (pt) (best single term)	0.93	0.92	0.49
	Therapy & (px)	0.95	0.62	0.15
	Therapeutic Use & (px)	0.89	0.70	0.18
	Therapeutic Use (sh)	0.70	0.84	0.24
	Drug Therapy (sh)	0.63	0.84	0.23
	Prevention & Control (sh)	0.26	0.91	0.16
	Therapy (sh)	0.14	0.90	0.08

For sensitivity, pre-explosions out-performed methodologic textwords and MeSH terms in 5 out of 8 instances. For treatment in 1991 and 1986 the single terms yielding the highest sensitivity, 'Clinical Trial (pt)' (93%) and 'Random: (tw)' (82%), were out-performed by 'Therapy& (px)' (95% for 1991 and 91% for 1986). For diagnosis in 1991 and 1986 the term yielding the highest sensitivity 'Sensitivity (tw)' (57% in 1991 and 43% in 1986) was out-performed by 'Diagnosis& (px)' (80% in 1991 and 79% in 1986). For etiology in 1986 the best single term 'Risk (tw)' (61%) was out-performed by 'Etiology& (px)' (68%). In all cases, however, use of the pre-explosions resulted in a loss in specificity with a corresponding loss in precision. For example, the 2% gain in sensitivity achieved when searching with 'Therapy& (px)' rather than 'Clinical Trial (pt)' in 1991 was coupled with a drop in specificity from 92% to 62% and a drop in precision from 49% to 15%.

Table 3. Sensitivity, specificity and precision for single best term and subheadings in 1986

Category	Search strategy	Sensitivity	Specificity	Precision
1986				
Etiology	Risk (tw) (best single term)	0.61	0.89	0.16
	Etiology& (px)	0.68	0.53	0.05
	Etiology (sh)	0.36	0.77	0.06
Prognosis	Prognosis (tw) (best single term)	0.56	0.97	0.29
	Mortality (sh)	0.44	0.95	0.18
Diagnosis	Sensitivity (tw) (best single term)	0.43	0.98	0.30
	Diagnosis& (px)	0.79	0.74	0.06
	Diagnosis (sh)	0.62	0.89	0.09
	Diagnostic use (sh)	0.16	0.96	0.10
Treatment	Random: (tw) (best single term)	0.82	0.95	0.53
	Therapy& (px)	0.91	0.62	0.13
	Therapeutic Use& (px)	0.83	0.64	0.13
	Drug Therapy (sh)	0.66	0.81	0.19
	Therapeutic Use (sh)	0.63	0.81	0.18
	Prevention & Control (sh)	0.16	0.94	0.12
	Therapy (sh)	0.13	0.90	0.07

DISCUSSION

Our findings show that in most instances, pre-explosions can achieve higher sensitivity for detecting sound clinical studies in MEDLINE than single methodologic textwords, subject headings, or

subheadings but at the cost of lower specificity and precision. These results were not found for prognosis and were inconsistent for etiology, suggesting that improvements in indexing are needed here.

It is worth noting that we had a pre-screening step in the development of our search strategies. When searching for each journal title in MEDLINE the publication types 'editorial', 'comment', 'letter', and 'news' were excluded from the search using the boolean 'AND NOT' operator. This pre-screening step would have no effect on the sensitivity calculated for the combinations of terms as studies meeting the key methodologic criterion were defined by the manual review of the literature. This step would, however, result in improvements of specificity and precision. Thus, searchers would be advised to include this pre-screening step if maintaining similar levels of specificity and precision are of concern.

The search strategies presented here can aid searchers, particularly clinicians who are inexperienced in constructing complex searches, to retrieve studies that meet at least one major criterion for scientific merit for applied health care research while filtering out studies with weaker designs. Such filters are bound to retrieve 'false positive' articles and miss others that should be retrieved. Retrieved articles must be further evaluated by the user to determine their methodologic soundness and clinical applicability. 'False negative' articles can only be retrieved by hand searching journals or other labor-intensive means.

Other possible quality filters such as ordering journals by impact factors and citations exist but we do not know how these methods compare with our search filters. However, even among the best journals only a small proportion of articles met the quality criteria we used.

One limitation of this study was that only priority journals were included in the search. Also, only the abstracts and titles of citations could be searched for textword inclusion. However, one of the strengths of this study was the highly reproducible classification of articles in the manual hand searches which served as the gold standard.

For most research purposes, we recommend that the search term with the highest sensitivity be used in the MEDLINE search so that key articles will not be missed. In back file searches the most appropriate term may differ and the search should be modified appropriately. For clinical searches, higher precision may be desirable especially if there is redundancy in the literature being retrieved.

Future research will have to address how these search terms perform when they are combined in all possible permutations and combinations of MeSH terms

and textwords.

Appendix. Complete List of Search Terms

Notes: Terms with 0 citations retrieved in 1986 are marked with *; terms with 0 citations retrieved in 1991 are marked with †; terms with < 10% sensitivity in 1991 are marked with ‡; terms with < 10% sensitivity in 1986 are marked with §; truncation is noted by :: the & indicates a subheading pre-explosion.

Etiology

MeSH terms: exp case control studies§; case control studies*; retrospective studies†§; exp cohort studies; cohort studies*; exp longitudinal studies; longitudinal studies†§; follow-up studies§; prospective studies; cross-sectional studies†§; exp causality*; causality*; risk factors*; exp risk; risk‡; logistic models*; odds ratio*; etiology& (px); etiology (sh); Textwords: cohort§; risk; etiol: or aetiol.; odds and ratio:§; causation and causal:‡; relative and risk; case and control.; case and comparison†§; case and referent*†.

Prognosis

MeSH terms: exp cohort studies; cohort studies*; exp longitudinal studies; longitudinal studies†§; follow-up studies; prospective studies; prognosis; exp morbidity§; morbidity†§; incidence*; exp mortality§; mortality†§; cause of death*†; infant mortality†§; maternal mortality†§; maternal mortality†§; survival rate*; survival analysis*; mortality (sh); Textwords: natural and history‡; prognos.; inception and cohort*†; clinical and course§; predict.; outcome.; clinical and consequence:‡§; prognostic and factor.; morbidity†§; course.

Diagnosis

MeSH terms: exp sensitivity and specificity§; sensitivity and specificity§; predictive value of tests§; ROC curve*†; exp diagnostic errors†§; diagnostic errors†§; false positive reactions†§; false negative reactions†§; diagnosis, differential†§; diagnosis& (px); diagnosis (sh); diagnostic use (sh); Textwords: sensitivity; specificity; predictive and value.; post and test and probabilit:†§; post and test and likelihood†§; likelihood and ratio:‡§; false and rate†§;

false and positive‡; false and negative†§; receiver and operat: and characteristic†§; roc†§; independent and comparison†§; mask: and comparison*†; blind: and comparison†§; gold and standard†§; pre and test and probability:*†; pre and test and likelihood*†; independent comparison*†.

Treatment

MeSH terms: exp research design; research design†§; double-blind method; random allocation‡; exp clinical trials*†; clinical trials‡; multicenter studies*†; randomized controlled trials*†; clinical trial (pt); exp multicenter studies*†; multicenter study (pt)*; randomized controlled trial (pt)*; comparative study; single-blind method*†; placebos†§; prevention & control (sh); therapy& (px); therapy (sh); drug therapy (sh); therapeutic use& (px); therapeutic use (sh); Textwords: random.; placebo.; double and blind.; mask:†§; single and blind:‡§; controlled and trial:.

ACKNOWLEDGEMENTS

The study was supported by the Ontario Ministry of Health and the National Library of Medicine (R01 LM04696-03).

References

- [1]. Haynes RB, McKibbin KA, Fitzgerald D, Guyatt GH, Walker CJ, Sackett DL. How to keep up with the medical literature. V. Access by personal computer to the medical literature. *Ann Intern Med* 1986;105:810-6.
- [2]. Wilczynski NL, Walker CJ, McKibbin KA, Haynes RB. Assessment of methodologic search filters in MEDLINE. *Proc Annu Symp Comput Appl Med Care*. 1994;17:601-5.
- [3]. Science Citation Index. Vol. 16: Journal Citation Reports, 1984. Philadelphia, Institute for Scientific Information; 1985.
- [4]. Sackett DL, Haynes RB, Guyatt GH, Tugwell P. *Clinical epidemiology: a basic science for clinical medicine*. Second Edition. Little, Brown and Company, Boston, 1991.